MATH 343 / 643 Homework #1

Professor Adam Kapelner

Due 11:59PM March 3, 2024

(this document last updated 11:46pm on Monday 19th February, 2024)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read as much as you can online about the topics we covered.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 7 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: _____

Problem 1

These are general questions about Gibbs Sampling and Metropolis-within-Gibbs Sampling.

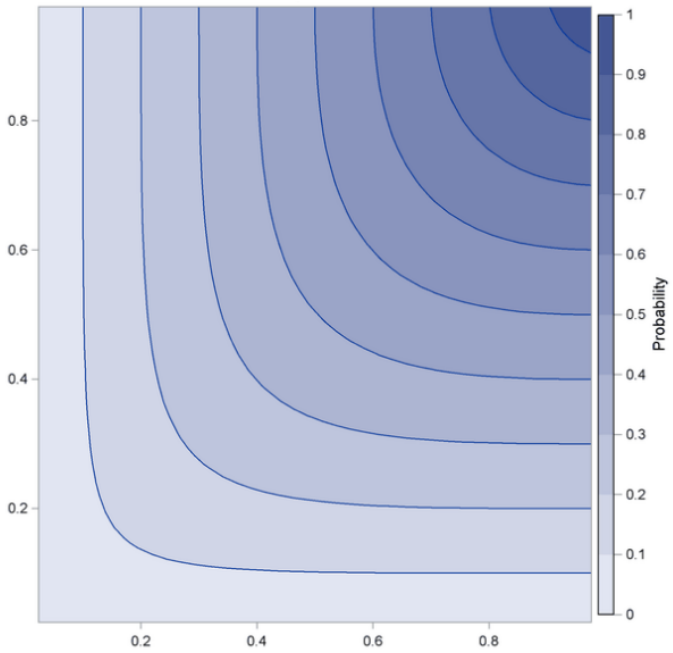
(a) [easy] Let $\dim[\boldsymbol{\theta}] = p$ and assume a prior $f(\boldsymbol{\theta})$ to be continuous. Describe the steps of the systematic sweep Gibbs Sampler algorithm below that will converge to $f(\boldsymbol{\theta} \mid \boldsymbol{X})$. Label the steps that are necessary for the p dimensions separately e.g. Step 2.1, Step 2.2, ..., Step 2.p. You need to reference these step numbers later on in the problem.

(b) [easy] What are all the items you need to know in order to write code for that implements a Gibbs Sampler?

(c) [easy] Explain what burning of the chain is and why it is necessary.

(d) [easy] Explain what thninng of the chain is and why it is necessary.

(e) [easy] Pretend you are estimating $\mathbb{P}(\theta_1,\ \theta_2 \mid X)$ and the joint posterior looks like the picture below where the x axis is θ_1 and the y axis is θ_2 and darker colors indicate higher probability. Begin at $[\theta_1, \theta_2] = [0.5, 0.5]$ and simulate 5 iterations of the systematic sweep Gibbs sampling algorithm by drawing new points on the plot.



(f) [easy] What are all the items you need to know in order to write code for that implements a Gibbs Sampler?

(g) [easy] What are all the items you need to know in order to write code for that implements a Gibbs Sampler?

(h) [easy] Consider the need to implement a Metropolis Hastings step within the Sampler for θ_j. Why would you need to do this? At which step (reference your steps in part a) would you require it?

(i) [easy] If $\text{Supp}[\theta_j] = \mathbb{R}$, propose a default proposal distribution to start with:

$q(\theta_{t,j} \,|\, \theta_{t-1,j}, \phi) =$

Remember, the mean of proposal distributions should be $\theta_{t-1,j}$ (or close to that value) and ϕ are additional parameters which may or may not be used.

(j) [harder] How do you know if this proposal distribution is a good choice or not?

(k) [difficult] If $\text{Supp}[\theta_j] = (0, \infty)$, propose a proposal distribution

$q(\theta_{t,j} \,|\, \theta_{t-1,j}, \phi) =$

(l) [difficult] [MA] If $\text{Supp}[\theta_j] = [0, 1]$, propose a proposal distribution

$q(\theta_{t,j} \,|\, \theta_{t-1,j}, \phi) =$

Problem 2

Consider a count model that has many zeroes. We choose to fit it with a hurdle model

$$X_1, \ldots, X_n \overset{iid}{\sim} \begin{cases} 0 & \text{w.p. } \theta_1 \\ \text{ShiftedExtNegBinomial}(\theta_2, \theta_3, +1) & \text{w.p. } 1 - \theta_1 \end{cases}$$

where the shifted distribution is just the extended negative binomial distribution so that the probability of realizing a count of one is the probability of realizing a count of zero, the probability of realizing a count of two is the probability of realizing a count of one, etc. i.e.

$$\text{ShiftedExtNegBinomial}(\theta_2, \theta_3, +1) := p(x) = \frac{\Gamma(x_i - 1 + \theta_2)}{(x_i - 1)!\Gamma(\theta_2)}(1 - \theta_3)^{x_i - 1}\theta_3^{\theta_2}.$$

(a) [harder] What is the parameter space for all three parameters of interest? This may require looking at your MATH 340 notes.

(b) [harder] Assume a flat prior $f(\theta_1, \theta_2, \theta_3) \propto 1$. Find the kernel of the posterior distribution $f(\theta_1, \theta_2, \theta_3 \,|\, \boldsymbol{x}, n_0, n_+)$ where $\boldsymbol{x} := \{x_1, \ldots, x_n\}$, the observations. Let n_0 be the number of zeroes in the dataset and $n_+ := n - n_0$, the number > 0 in the dataset.

(c) [harder] Find the log of the kernel of the posterior distribution.

(d) [easy] Find the conditional distribution $f(\theta_1 \,|\, \boldsymbol{x}, n_0, n_+, \theta_2, \theta_3)$ as a brand name rv.

(e) [easy] Find the kernel of the conditional distribution $f(\theta_2 \,|\, \boldsymbol{x}, n_0, n_+, \theta_1, \theta_3)$.

(f) [easy] Is the conditional distribution $f(\theta_2 \,|\, \boldsymbol{x}, n_0, n_+, \theta_1, \theta_3)$ a brand name rv? Yes/no

(g) [easy] Given your answer in (a), the Supp $[\theta_2]$ and your answer from problem 1(k) which was marked difficult, provide a proposal distribution

$q(\theta_{t,2} \,|\, \theta_{t-1,2}, \phi) =$

(h) [easy] Find the conditional distribution $f(\theta_3 \,|\, \boldsymbol{x}, n_0, n_+, \theta_1, \theta_2)$ as a brand name rv.

Problem 3

These are general questions about Permutation Testing.

(a) [easy] What are the null and alternative hypotheses for a two-sample permutation test?

(b) [easy] Let n_1 and n_2 be the sample sizes from population one and population two respectively. How many possible sample "permutations" are there? I put permutations in quotes because it's not truly a "permutation" in the sense that you were taught in MATH 241.

(c) [easy] Give three examples of a test statistic to employ within the body of the loop of a permutation test.

(d) [difficult] Explain how you would calculate a p-value in a permutation test.

Problem 4

These are general questions about the Bootstrap. Assume $X_1, \ldots, X_n \overset{iid}{\sim}$ some DGP.

(a) [easy] Describe the steps in the bootstrap procedure for the estimate $\hat{\hat{\theta}} := w(x_1, \ldots, x_n)$ which estimates θ.

(b) [easy] In what situations should the bootstrap be employed instead of other inferential procedures you learned about?

(c) [difficult] Explain in what situations the bootstrap fails.

Problem 5

These are questions about parametric survival using the Weibull model i.e.

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Weibull}(k, \lambda) := f(y) = k\lambda^k y^{k-1} e^{-\lambda^k y^k} \mathbb{1}_{y>0}, \quad F(y) = 1 - e^{-\lambda^k y^k}, \quad S(y) = e^{-\lambda^k y^k}$$

(a) [difficult] Assume no censoring in the data. Find closed form expressions and/or equations for the MLEs of k and λ

(b) [difficult] Assume censoring in the data so that $\mathbf{c}$ is the binary vector that is one when censored and zero if measured. Let $\boldsymbol{y}$ be the vector of measurements or censored values if not measured. Find $\ell(k, \lambda; \boldsymbol{y}, \boldsymbol{c})$.

(c) [harder] In class we proved that $\mathbb{E}[Y] = \frac{1}{\lambda}\Gamma\left(1+\frac{1}{k}\right)$. Use this result to find $\mathbb{E}[Y \mid Y > a]$ where $a > 0$. You should first find the density of the truncated distribution. Then the expectation of this distribution will basically follow the same steps as found in lecture when we derived the expectation.

(d) [harder] Describe the steps in an EM algorithm to find the maximum likelihood estimates of k and λ.

Problem 6

These are questions about nonparametric survival inference.

(a) [harder] Explain how the Kaplan-Meier estimator differs from the empirical survival function if there is censoring at all different times before and after the maximum measured survival. There is only one difference!

(b) [easy] Consider the dataset $y = \{79, 81, 92+, 95, 105+, 107, 122\}$ where the "+" signs indicate censored values. Draw the Kaplan-Meier estimate of $S(y)$. Try to make it to scale as best as possible.

(c) [easy] Write the hypotheses for the log-rank test.

(d) [easy] Write the formula for the test statistic in the log-rank test.